缺失值处理方法探索：基于个体水平数据的卫生经济学评价*

宋若萌[①]，郝 军[②③④]，云 科[⑤]，李汶檀[①]，章溪妍[①]，辛 雨[⑥]，吴昌金[①]，蔡源益[①]，吴华章[①]，惠 文[⑥]

摘 要 目的：文章系统探讨基于个体水平数据的卫生经济学评价中缺失值的系列问题，为实际研究中正确处理和报告缺失值提供参考。方法：在回顾一般缺失值问题的基础上，梳理基于个体水平数据的卫生经济学评价中数据缺失的原因、类型和处理方法等。结果：多重插补法是缺失值处理方法最常见的方法，在使用该方法时要注意合理选择插补具体方法、插补建模类型和插补模型的变量。结论：当前卫生经济学评价中关于缺失值的处理和报告还未形成已达成共识的质量规范，有待未来进一步探索。

关键词 统计数据缺失；个体水平数据；卫生经济学评价

中图分类号 R1-9；R-012 文献标志码 B 文章编号 1003-0743(2023)07-0006-04

Exploring the Missing Data in Health Economic Evaluations: Health Economic Evaluation Based on Individual Level Data/Song Ruomeng, Hao Jun, Yun Ke, et al.//Chinese Health Economics, 2023,42(7):6-9

Abstract Objective: To systematically explore a series of issues related to missing data in health economics evaluation based on individual level data, providing references for correctly handling missing data in practical research. **Methods:** Based on reviewing the basic of missing data, it summarized the causes, types, and handling methods of missing data in health economics evaluation based on individual level data. **Results:** Multiple-imputation is the most common method for handling missing data. In the application of the method, care should be taken to choose the specific method of interpolation, the type of interpolation modelling and the variables of the interpolation model wisely. **Conclusion:** There is no universal agreement quality standard for the handling and reporting of missing data in health economic evaluation, which needs to be further explored in the future.

Keywords missing statistical data; individual level data; health economic evaluation

First-author's address School of Health Management, China Medical University, Shenyang, 110122, China

Corresponding author Hui Wen, E-mail: huiwen@wchscu.cn

卫生经济学评价（Health Economic Evaluation，HEE）是对多个可替代医疗方案的成本和效果的比较分析[1]，一般可分为基于模型的HEE和基于个体水平数据的HEE[2]。其中，基于个体水平数据的HEE是指在传统随机对照试验和真实世界研究中加载开展，该类HEE常常由于失访、随访记录不完全、资源使用数据不可得等原因导致成本和效果数据缺失。据报道，HEE中仅有60%~75%的参与者具有完整的成本和效果数据[3]。数据缺失会影响统计效力，产生偏倚效应[4]。当缺失值占比很大时，甚至可能会对HEE结果产生颠覆性影响[5]。因此，在基于个体水平数据的HEE中合理处理缺失数据是提高研究质量的重要方法之一。

目前，我国关于临床研究中一般缺失值的缺失原因、类型和处理方法等方面已有探讨，但对基于个体水平数据的HEE缺失值问题尚未开展系统研究。因此，本研究在回顾一般缺失值的基本理论和处理方法基础上，进一步探讨个体水平数据的HEE缺失值处理问题，为我国相关研究提供方法学参考。

1 缺失值的概述

1.1 定义及分类

缺失值是缺失的拟获取的观测值[6]。其分为两种情况，其一，客观存在的缺失值，也称系统缺失，此时不需要对其处理；其二，主观判定的缺失值，即经研究者判断认为应该得到但实际未得到的数据，在此情况下，缺失值>5%时需要对其进行处理[7]。

1.2 常见缺失原因

数据缺失的原因主要归为3点：一是问题不适用、技术无法获取。二是未获取有效信息或获取的数据不可用。三是无回答，包括单元无回答和条目无回答。其中，单元无回答是指由于个体失访、退出等原因未接受调查，使得个体的所有数据均缺失；条目无回答是个体未有效回答全部问题，部分数据缺失。

* 基金项目：教育部人文社科基金（22YJCZH065）；四川省自然科学基金（2023NSFSC1046）。

① 中国医科大学健康管理学院 沈阳 110122
② 中国医学科学院北京协和医学院 北京 102300
③ 中国医学科学院阜外医院 北京 102300
④ 国家心血管病中心医学统计部 北京 102300
⑤ 中国医科大学附属第一医院 沈阳 110122
⑥ 四川大学华西医院 成都 610041

作者简介：宋若萌（1997—），女，博士在读；研究方向：卫生经济学评价及其循证评价；E-mail: 1044223732@qq.com。

通信作者：惠文，E-mail: huiwen@wchscu.cn。

1.3　缺失值的主要类型

缺失类型有4种：第一，根据包含缺失值的变量个数，分为单变量缺失和多变量缺失。第二，根据缺失模式，分为单调缺失和非单调缺失。第三，根据缺失机制，分为完全随机缺失（Missing Completely at Random, MCAR）、随机缺失（Missing at Random, MAR）和非随机缺失（Missing Not at Random, MNAR）[8]。其中，MCAR是指缺失数据的概率与已观测到的数据和未观测到的数据均无关。MAR是指在一定的条件下，缺失数据的概率与未观测到的数据无关，而与部分已观测到的数据有关。MNAR是指缺失数据的概率与未观测到的数据有关。第四，根据缺失数据的影响，分为可忽略的缺失和不可忽略的缺失[7]。

1.4　处理方法

缺失值的处理方法主要分为加权法、删除法和插补法。加权法适用于单元无回答，即将缺失单元的权数分解到非缺失单元上。删除法是将存在缺失值的个体删除，根据删除的程度分为完全案例分析（Complete Case Analysis, CCA）和有效案例分析。插补法是研究最深入、最广泛的方法，是用估计值插补缺失值的方法，包括单一插补和多重插补（Multiple Imputation, MI）。单一插补是通过统计方法获取一个估计值来插补缺失值，包括均值插补、回归插补、热平台插补、冷平台插补、末次观测结转法（Last Observation Carried Forward, LOCF）等。MI是构造多个估计值，再综合为一个最佳参数估计值进行插补。

2　HEE中的缺失值问题

2.1　缺失原因

HEE中数据缺失的原因除常见的原因，又有特殊性。首先，HEE一般使用终点指标，而非中间指标作为效果指标，故随访周期通常较长，患者易失访，从而出现缺失值。其次，HEE研究视角对应不同的成本类型，越宽泛和全面的研究视角所需要的成本参数越多，越易产生缺失值。最后，质量调整生命年是最常用的效果指标。当通过患者自报告的方式获取该指标值时，量表可能丢失或不完整，导致数据缺失[9]。

2.2　缺失类型

除前文所述按照缺失模式、缺失机制的分类方式外，HEE还可以根据缺失数据的性质分为基线数据缺失、成本数据缺失和效果数据缺失[10]。不同性质的缺失数据需要单独处理，如果数据缺失的特点不同，所选择的处理方法也可能不同。

2.3　处理方法

当数据缺失程度＜50%时，采取恰当的方法可以减少结论的偏差[9]。目前并没有适用于一切情景的达成共识方法，应根据缺失数据的特点（纵向结构、非正态分布、相关性）、缺失类型及样本量等因素，选择适合不同研究设计的方法[10]。以下重点介绍删除法和插补法在HEE中的应用。

2.3.1　删除法。HEE中常见的删除法是CCA。CCA假设拥有完整数据的个体可以代表存在缺失数据的个体，故而仅保留并分析能够观测到全部成本和效果数据的个体。如表1所示，仅第5个个体纳入分析范围。该方法适用于MCAR下样本量很大、缺失值不多（条目无回答＜25%或单元无回答＜10%[11]）且完全观察的样本比例≥70%的情况。CCA处理简便，但局限性也很突出。CCA删除了大量可能有效的信息，会造成信息浪费；如果缺失数据的真实值与现有数据偏差较大，则会影响结论的有效性。

表1　CCA示意

个体	成本数据		效果数据
	医疗资源X_1使用量A	X_1的单价X_2	量表条目X_3
1	数据完整	数据缺失	数据完整
2	数据完整	数据完整	数据缺失
3	数据缺失	数据完整	数据缺失
4	数据缺失	数据缺失	数据完整
5	数据完整	数据完整	数据完整

2.3.2　插补法。

（1）单一插补。HEE中常见的单一插补法为均值插补和LOCF。均值插补是指用已观测值的平均值填补缺失数据。均值插补适用于MCAR，要求数据缺失不超过10%[12]，是基线数据缺失时最有效的方式。其局限性在于：低估方差或标准误；由于未考虑其他协变量的信息，淡化数据结构；不适用于处理缺失的效果数据[10]。公式为：

$$\bar{y}=\frac{\sum_{i=1}^{n}a_i y_i}{n} \qquad 式（1）$$

$\bar{y}$表示插补值，n有观测值的单元数，有观测值时$a_i=1$，缺失数据时$a_i=0$，y为观测值。

LOCF假设最后一次观测值是后续观测中缺失数据的代表，可用于效果数据的缺失。如研究中随访有多个时间点T_1、T_2、T_3等，当T_2的效果数据缺失时，则以T_1的观测值为插补值填补于T_2的观测值中。其缺点是这种方法难以控制Ⅰ类错误，会影响检验效能和估计误差[13]。

（2）MI。MI是HEE中处理缺失值的主流方法，可用于基线、成本和效果数据缺失。MI适用于缺失机制为MCAR、MAR和MNAR的情况[14]，其中以MAR假设最为常见。HEE的MI遵循3个步骤：一是插补。针对一个缺失值构造多个插补值，从而形成多个完整数据集。二是分析。对多个完整数据集进行回归分析，产生多个参数估计。三是综合。合并多个分析结果，得到最终参数估计，见图1。

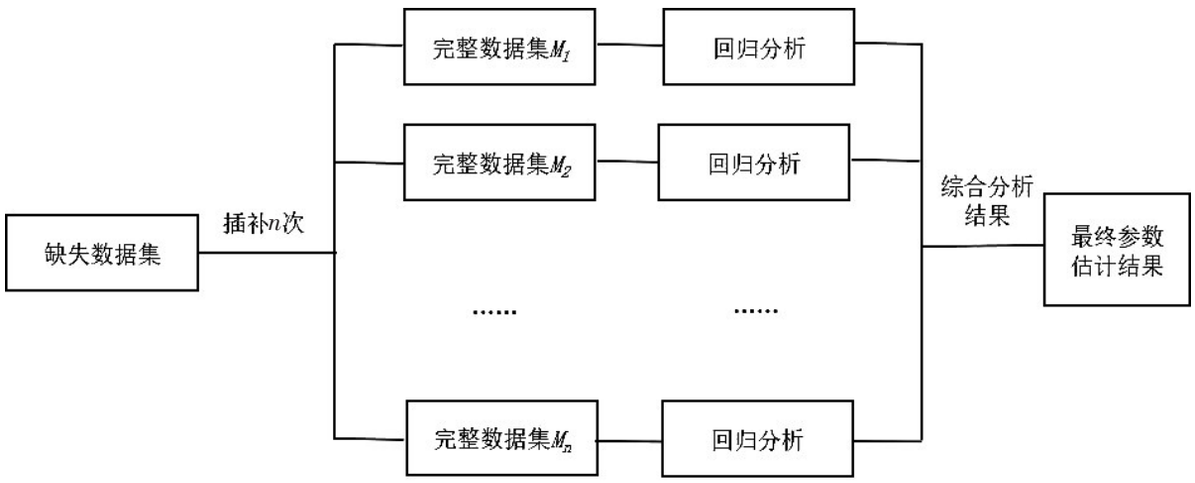

图1　多重插补原理

MI具体包括多种方法：一是马尔可夫链蒙特卡罗法（Markov Chain Monte Carlo，MCMC）。MCMC是一种基于贝叶斯理论求解后验分布的方法，重复循环n次计算填补和后验两个步骤，得到n个完整的数据集。在HEE的整个周期中，当成本和效果的数据缺失是由于中间某个时间点的单元无回答或者是某个条目无回答导致的，此时为非单调缺失，优选MCMC法。二是回归预测法。以带有缺失值的某一变量为因变量，以无缺失值且与缺失数据相关的变量为辅助变量，建立回归模型，并根据得到的模型插补缺失值[7]。三是预测均值匹配法（Predictive Mean Matching Method，PMM）。PMM属于回归预测法的一种变形，它通过线性回归模型得到回归系数的参数估计，用最接近预测均值的实际观测值来替代缺失值[7]。该方法适用于成本和效果数据是非正态分布的填补。

实施MI主要有两种建模方式：联合建模法（Joint Modeling，JM）和链式方程法（Multivariate Imputation using Chained Equations，MICE）。成本和效果的缺失数据常是多变量任意缺失模式，当符合正态分布情况时，优选JM法；当不符合正态分布时，选择MICE法。由于成本和效果数据一般是非正态分布的，故MICE是实践中更为流行的建模方式。此外，当缺失的数据是纵向性质时，在进行MI时优选JM，以减少对缺失数据复合填补的偏差，但需要先将数据转化为正态分布[15]。

插补模型中应包括与插补值相关的所有变量[16]。在实际操作中，尽管原则上需要纳入的变量很多，但为了保证模型的精准性和有效性，应选择最具预测性的变量纳入其中。插补模型建立后需要验证，常用的验证方法有两种，一是评估观测值和预测值的分布是否相似；二是将通过插补模型获得的结果与缺失值处理替代方法获得的结果相比较[10,14]。值得注意的是，当成本和效果均采取MI时，插补方法、建模方式以及插补模型中的变量可能不同。

2.4　注意事项

2.4.1　条目或总体层级处理。当HEE中的成本和效果缺失的数据属于复合指标，且缺失程度＞20%时，在条目层级或总体层级进行处理可能会得到不同的结论[17]。

对于效果指标，当采用患者自报告的方式获取时，可能发生单元无回答或条目无回答。如EQ-5D量表，总分由5个条目的分项得分构成，即行动能力、日常活动、自我护理、疼痛和不适以及焦虑和抑郁。当样本量＞500个时，两种层级插补结果相似；若仅存在条目无回答或问卷内数据的缺失程度≥20%时，条目层级的处理更准确；当样本量＜100个时，优选对条目进行处理[14]。对于成本指标，当用于计算总成本的各类资源使用量具有相同缺失机制时，在总体层级处理缺失值更合适；当用于计算总成本的各类资源消耗量具有不同缺失机制时，需要在条目层级处理缺失值[18]。

MI是处理HEE中缺失值的优选，要根据数据缺失的程度和插补方式选择合适的多重插补模型和方法。当对条目层级进行插补时，推荐使用基于JM的回归预测和PMM法；当对总体层级进行插补时，使用PMM法和MCMC法更佳[19]。

2.4.2　敏感性分析。在数据插补后需要进行敏感性分析。敏感性分析可以比较不同缺失值处理方法对结果的影响、验证缺失机制和校正插补方法等[10,18]。在缺失机制的敏感性分析中，主分析中通常假定缺失机制为MAR，为验证结论在非MAR假设下的稳健性，敏感性分析中要假定缺失机制为MNAR。在插补方法的敏感性分析中，需要根据变量的类型选择与主分析不同的方法。例如当主分析采用CCA时，敏感性分析可采用LOCF；当主分析采用MI时，敏感性分析可采用CCA；纵向数据的缺失还可采取贝叶斯法[20]。当敏感性分析结

果与主分析结果产生相似结论时，说明缺失值的处理方法不会对整体结论产生重要的影响；当敏感性分析的结果与主分析的结果不一致时，需要报告并分析其原因[6]。

3　讨论

在基于个体水平数据的HEE中，数据缺失是非常普遍的情况，严重的数据缺失会导致研究结论缺乏真实性和可靠性，形成的证据难以支撑决策[17]。因此，在研究过程中首先要尽可能避免数据缺失。首先，针对数据缺失的可能原因，在研究设计、实施阶段就采取相应控制策略，同时，根据预测缺失比例合理设定样本量大小，以保证统计效力[4]。其次，当发生需要处理的数据缺失时，应根据缺失数据的特点、类型和缺失机制等，选择合理的处理方法。最后，采用敏感性分析验证缺失值处理的稳健性。目前，国外学者针对临床试验中缺失值的处理制定了相应的方法学规范和报告规范[4,6]，但对于HEE缺失值处理的规范要求，仅在ISPOR实践规范中提及“在进行经济分析时，需要对缺失数据做出解释，缺失值的数量、处理缺失值的方法、分析缺失值处理方法的不确定性等信息均需报告”[21]。近年来，尽管有学者提出制定HEE缺失值处理的建议[10]，但至今该领域仍缺乏权威的方法学、报告或指南，这有待今后开展更为全面和深入的研究。

参　考　文　献

[1] DRUMMOND M F, SCULPHER M J, KARL C, et al. Methods for the economic evaluation of health care programmes[M]. New York: Oxford University Press, 2015.

[2] 刘国恩，胡善联，吴久鸿，等. 2020中国药物经济学评价指南（中英双语版）[M]. 北京：中国市场出版社，2020.

[3] GABRIO A, PLUMPTON C, BANERJEE S, et al. Linear mixed models to handle missing at random data in trial-based economic evaluations[J]. Health econ, 2022,31(6): 1276-1287.

[4] Committee for Medicinal Products for Human Use. Guideline on missing data in confirmatory clinical trials[EB/OL]. (2010-07-022) [2023-03-04]. http://www.cajcd.edu.cn.html

[5] RODERICK J A L, DONALD B R. Statistical analysis with missing data[M]. Hoboken: John Wiley & Sons, 2002.

[6] CARPENTER J R, KENWARD M G. Missing data in randomised controlled trials-a practical guide[M]. Birmingham: Health Technology Assessment Methodology Programme, 2007.

[7] 严杰. 缺失数据的多重插补[M]. 重庆：重庆大学出版社，2017.

[8] RUBIN D B. Inference and missing data[M]. Biometrika: Oxford University Press, 1976.

[9] BEN A J, VAN DONGEN J M, ALILI M E, et al. The handling of missing data in trial-based economic evaluations: should data be multiply imputed prior to longitudinal linear mixed-model analyses?[J]. Eur j health econ, 2022, Online ahead of print.

[10] FARIA R, GOMES M, EPSTEIN D, et al. A guide to handling missing data in cost-effectiveness analysis conducted within randomised controlled trials[J]. Pharmacoeconomics, 2014,32(12): 1157-1170.

[11] BOWRIN K, BRIERE J B, LEVY P, et al. Cost-effectiveness analyses using real-world data: an overview of the literature[J]. J med econ, 2019,22(6):545-553.

[12] EEKHOUT I, DE VET H C, TWISK J W, et al. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level[J]. J clin epidemiol, 2014, 67(3):335-342.

[13] 陈丽嫦，衡明莉，王骏，等. 定量纵向数据缺失值处理方法的模拟比较研究[J]. 中国卫生统计，2020, 37(3): 384-388.

[14] SIMONS C L, RIVERO-ARIAS O, YU L M, et al. Multiple imputation to deal with missing EQ-5D-3L data: Should we impute individual domains or the actual index?[J]. Qual life res, 2015,24(4):805-815.

[15] GABRIO A, HUNTER R, MASON A J, et al. Joint longitudinal models for dealing with missing at random data in trial-based economic evaluations[J]. Value health, 2021,24(5): 699-706.

[16] WHITE I R, ROYSTON P,WOOD A M. Multiple imputation using chained equations: Issues and guidance for practice[J]. Stat Med, 2011,30(4):377-399.

[17] MICHALOWSKY B, HOFFMANN W, KENNEDY K, et al. Is the whole larger than the sum of its parts? Impact of missing data imputation in economic evaluation conducted alongside randomized controlled trials[J]. Eur j health econ, 2020,21(5):717-728.

[18] LING X, GABRIO A, MASON A, et al. A scoping review of item-level missing data in within-trial cost-effectiveness analysis[J]. Value health, 2022,25(9):1654-1662.

[19] NOORAEE N, MOLENBERGHS G, ORMEL J, et al. Strategies for handling missing data in longitudinal studies with questionnaires[J]. Journal of statistical computation and simulation, 2018,88(17):3415-3436.

[20] MASON A J, GOMES M, CARPENTER J, et al. Flexible bayesian longitudinal models for cost-effectiveness analyses with informative missing data[J]. Health econ, 2021,30(12): 3138-3158.

[21] RAMSEY S, WILLKE R, BRIGGS A, et al. Good research practices for cost-effectiveness analysis alongside clinical trials: the ISPOR RCT-CEA task force report[J]. Value health, 2005,8(5):521-533.

[收稿日期：2023-04-27]（编辑：高非）